Pg. 220—226, #2, 4, 5, 9, 11, 22

2. Ages of couples.

a) The correlation between age difference and year is $r = \sqrt{R^2} = \sqrt{0.716} \approx -0.846$. The negative value is used since the scatterplot shows that the association is negative, strong, and linear.

b) The linear regression model that predicts age difference from year is: $(Men - \hat{Women}) = 33.483 - 0.015756(Year)$. This model predicts that each passing year is associated with a decrease of approximately 0.016 years in the difference between male and female marriage age. A more meaningful comparison might be to say that the model predicts a decrease of approximately 0.16 years in the age difference for every 10 years that pass.

c)

$(Men - \hat{Women}) = 33.483 - 0.015756(Year)$

$(Men - \hat{Women}) = 33.483 - 0.015756(2010)$

$(Men - \hat{Women}) \approx 1.81344$

According to the model, the age difference between men and women at first marriage is expected to be approximately 1.81 years. (This figure is very sensitive to the number of decimal places used in the model.)

d) The latest data point is before the year 2000. Extrapolating for 2010 is risky because it depends on the assumption that the trend in age at first marriage will continue in the same manner.

4. Ages of couples, again.

a) The data from the late 1800s to 1950 are high leverage points. Since they generally follow the same linear trend as the 1975 – 1998 data, those data points increase the correlation and the R^2 value.

b) The residuals plot shows no apparent pattern, so the linear model is appropriate.

c) For every 10 years that pass, the model predicts a decrease of approximately 0.24 years in average age difference at first marriage.

d) The y-intercept is the prediction of the model in year 0, over 2000 years ago. An extrapolation that far into the past is not meaningful. The earliest year for which we have data is 1975.

5. Good model?

a) The student's reasoning is not correct. A scattered residuals plot, not high R^2, is the indicator of an appropriate model. Once the model is deemed appropriate, R^2 is used as a measure of the strength of the model.

b) The model may not allow the student to make accurate predictions. The data may be curved, in which case the linear model would not fit well.

Pg. 220—226, #2, 4, 5, 9, 11, 22

9. Heating.

a) The model predicts a decrease in $2.13 in heating cost for an increase in temperature of 1° Fahrenheit. Generally, warmer months are associated with lower heating costs.

b) When the temperature is 0° Fahrenheit, the model predicts a monthly heating cost of $133.

c) When the temperature is around 32° Fahrenheit, the predictions are generally too high. The residuals are negative, indicating that the actual values are lower than the predicted values.

d)

$\hat{C} = 133 - 2.13(Temp)$

$\hat{C} = 133 - 2.13(10)$

$\hat{C} = \$111.70$

According to the model, the heating cost in a month with average daily temperature 10° Fahrenheit is expected to be $111.70.

e) The residual for a 10° day is approximately –$6, meaning that the actual cost was $6 less than predicted, or $111.70 – $6 = $105.70.

f) The model is not appropriate. The residuals plot shows a definite curved pattern. The association between monthly heating cost and average daily temperature is not linear.

g) A change of scale from Fahrenheit to Celsius would not affect the relationship. Associations between quantitative variables are the same, no matter what the units.

11. Unusual points.

a) **1)** The point has high leverage and a small residual.
2) The point is not influential. It has the *potential* to be influential, because its position far from the mean of the explanatory variable gives it high leverage. However, the point is not *exerting* much influence, because it reinforces the association.
3) If the point were removed, the correlation would become weaker. The point heavily reinforces the positive association. Removing it would weaken the association.
4) The slope would remain roughly the same, since the point is not influential.

b) **1)** The point has high leverage and probably has a small residual.
2) The point is influential. The point alone gives the scatterplot the appearance of an overall negative direction, when the points are actually fairly scattered.
3) If the point were removed, the correlation would become weaker. Without the point, there would be very little evidence of linear association.

4) The slope would increase, from a negative slope to a slope near 0. Without the point, the slope of the regression line would be nearly flat.

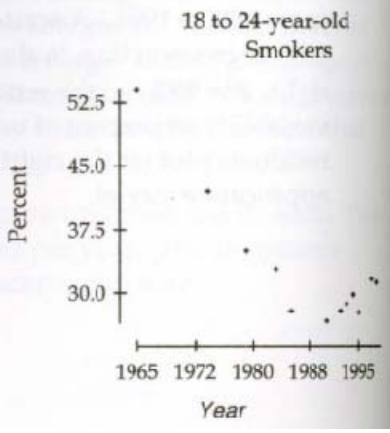
c) 1) The point has moderate leverage and a large residual.
2) The point is somewhat influential. It is well away from the mean of the explanatory variable, and has enough leverage to change the slope of the regression line, but only slightly.
3) If the point were removed, the correlation would become stronger. Without the point, the positive association would be reinforced.
4) The slope would increase slightly, becoming steeper after the removal of the point. The regression line would follow the general cloud of points more closely.

d) 1) The point has little leverage and a large residual.
2) The point is not influential. It is very close to the mean of the explanatory variable, and the regression line is anchored at the point $(\bar{x}, \bar{y})$, and would only pivot if it were possible to minimize the sum of the squared residuals. No amount of pivoting will reduce the residual for the stray point, so the slope would not change.
3) If the point were removed, the correlation would become slightly stronger, decreasing to become more negative. The point detracts from the overall pattern, and its removal would reinforce the association.
4) The slope would remain roughly the same. Since the point is not influential, its removal would not affect the slope.

22. Smoking.

The analysis that follows is one of several good models that may be used to predict the percentage of smokers among males ages 18 to 24. The important feature to recognize is that these data consist of two distinct trends. Your modeling decisions may vary slightly from these, but that is fine as long as those decisions are justified.

A scatterplot (at the right) of year vs. percent of males ages 18 to 24 who smoke shows two distinct trends. From 1965 to 1985, there is a strong, negative linear association between year and percent smokers. As time passed, the percentage of smokers decreased. For the years 1990 to 1998, there is a reasonably strong, positive, linear association between year and percent smokers. The percentage of smokers increased during this time period. Two linear models, used together, will fit the relationship well.



Pg. 220—226, #2, 4, 5, 9, 11, 22

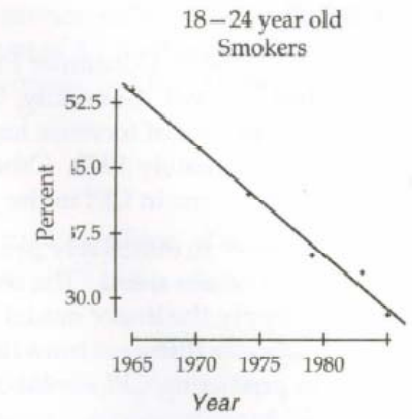
Model I (1965—1985)

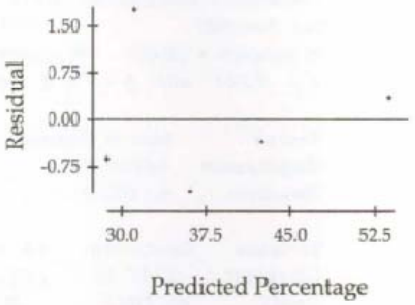
Dependent variable is: **Percentage**
No Selector
R squared = 98.8% R squared (adjusted) = 98.3%
s = 1.302 with 5 - 2 = 3 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	405.059	1	405.059	239
Residual	5.08900	3	1.69633	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	2521.62	160.7	15.7	0.0006
Year	-1.25592	0.0813	-15.5	0.0006

$\hat{\%} = 2521.62 - 1.25592(Year)$ is a good model for the years 1965 to 1985. A scatterplot of the relationship, with regression line, is shown at the right. $R^2 = 98.8\%$, so the model explains 98.8% of the variability in the percentage of males 18 - 24 years old who smoke. The residuals plot is scattered indicating an appropriate model.





odel II (1990—1998)

Dependent variable is: **Percentage**
No Selector
R squared = 77.1% R squared (adjusted) = 72.6%
s = 0.9874 with 7 - 2 = 5 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	16.4427	1	16.4427	16.9
Residual	4.87442	5	0.974884	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	-1152.14	287.6	-4.01	0.0103
Year	0.592378	0.1442	4.11	0.0093

$\hat{\%} = -1152.14 + 0.592378(year)$ is a decent model for the years 1990—1998. $R^2 = 77.1\%$, so only 77.1% of the variability in percent smokers is explained by the model. The residuals plot is acceptable, but might have a bit of a pattern. Removing points for 1995 and 1998 would increase R^2 and leave scattered residuals, but removing these points is impossible to justify. Removing 2 of 7 data points for no reason (other than the fact that they don't seem to fit the pattern we *think* is present) just isn't a good idea. The model given for these years seems to be the best we can do.